Heidrun Wiesenmüller, Magnus Pfeffer

Abgleichen, anreichern, verknüpfen

Das Clustering-Verfahren – eine neue Möglichkeit für die Analyse und Verbesserung von Katalogdaten

Neben den gewohnten Vortragsveranstaltungen in großen Sälen wartete der Leipziger Bibliothekskongress im März 2013 mit einem neuen Veranstaltungsformat auf: Verschiedene Workshops boten die Gelegenheit, Themen intensiv zu beleuchten und in kleinen Gruppen zu diskutieren. Einer dieser Workshops wurde von den Autoren des vorliegenden Beitrags gestaltet und war neuartigen Möglichkeiten für die Analyse und Verbesserung von Katalogdaten gewidmet. Als dritter Referent wurde Markus Geipel von der Deutschen Nationalbibliothek (DNB) über Google Hangout virtuell zugeschaltet. Initiiert wurde die Veranstaltung von der AG Bibliotheken der Deutschen Gesellschaft für Klassifikation, die damit an ihre Hildesheimer Tagung von 2012 anknüpfte.[1] Im Folgenden werden die wichtigsten Ergebnisse zusammengefasst.

Ein vergleichsweise einfaches Verfahren bildet die Grundlage: Über einen Abgleich einiger weniger Kategorien lassen sich mit großer Zuverlässigkeit diejenigen bibliografischen Datensätze aus einem Datenpool (der auch aus mehreren Katalogen bestehen kann) zusammenführen, die zum selben Werk gehören. Ein solches Werk-Cluster umfasst dann unterschiedliche Ausgaben und Auflagen eines Werkes ebenso wie Übersetzungen. Zu einem Cluster gehören alle Datensätze, die im Einheitssachtitel beziehungsweise in Sachtitel und Zusätzen übereinstimmen und mindestens eine verknüpfte Person oder Körperschaft gemeinsam haben.

Anreicherung von Sacherschließungsinformation

Das beschriebene Verfahren wurde von Magnus Pfeffer entwickelt und erstmals im Jahr 2010 erprobt. Es kam zunächst im Bereich der Sacherschließung zur Anwendung. Denn in Verbundkatalogen sind nicht selten mehrere Ausgaben desselben Werkes ganz unterschiedlich erschlossen: Ein Teil der Datensätze ist mit Schlagwörtern versehen, ein anderer mit Notationen aus unterschiedlichen Klassifikationen und wieder andere Datensätze haben gar keine inhaltliche Erschließung. Mithilfe des Werk-Clustering lässt sich nun die an verschiedenen Stellen vorhandene Sacherschließung sozusagen »poolen«: Die nicht oder nur zum Teil sachlich erschlossenen Datensätze können mit Erschließungsinformationen aus anderen Datensätzen desselben Clusters angereichert werden.

2011 wurde das Verfahren anhand der Daten des Südwestdeutschen Bibliotheksverbunds (SWB) und des Hessischen Bibliotheks- und Informationssystems (HeBIS) erfolgreich getestet. Das Einspielen der innerhalb eines Clusters gefundenen Schlagwörter und Notationen der Regensburger Verbundklassifikation (RVK) für alle Datensätze führte zu einem nicht unerheblichen Anwachsen der Sacherschließungsrate in beiden Verbünden. Ein Jahr später wurden auch die Daten des B3Kat – des gemeinsamen Verbundkatalogs von Bibliotheksverbund Bayern und Kooperativem Bibliotheksverbund Berlin-Brandenburg – sowie des Hochschulbibliothekszentrums Nordrhein-Westfalen (hbz) in den Abgleich mit einbezogen. Der Zuwachs an Sacherschließung wurde dadurch nochmals erhöht. Abbildung 1 (diese Seite) zeigt die mögliche Zunahme von RSWK- und RVK-Erschließung nach

> Für zwei unterschiedliche Erschließungssysteme werden nur die Cluster betrachtet, die Notationen aus beiden Systemen enthalten.

Katalog	Datensätze Monografien	Anteil RVK	Anteil RSWK	Zuwachs RVK	Zuwachs RSWK
SWB	13 330 743	4 217 226	4 083 113	581 780	957 275
HeBIS	8 844 188	1 933 081	2 237 659	1 097 992	1 308 581
hbz	13 271 840	1 018 298	3 322 100	2 272 558	1 080 162
B3Kat	22 685 738	5 750 295	6 055 164	2 969 381	2 765 967

Abbildung 1: Ergebnisse der Anreicherung aufgrund des Werk-Clustering

1 Vergleiche den Vortrag von Magnus Pfeffer im Rahmen der Jahrestagung der Deutschen Gesellschaft für Klassifikation 2012, http://digbib.ubka.uni-karlsruhe.de/volltexte/1000029147, und Magnus Pfeffer (2013): Using clustering across union catalogues to enrich entries with indexing information. In: Data analysis, machine learning and knowledge discovery: proceedings of the 36th Annual Conference of the Gesellschaft für Klassifikation e.V. in Hildesheim, Germany. Berlin, Heidelberg: Springer (im Druck)

2 Im SWB waren bereits RVK-Daten aus der ersten Projektphase 2010 eingespielt worden, wodurch der Zuwachs nicht mehr so groß ausfiel wie bei den anderen Katalogen.

3 Vergleiche dazu den Vortrag von Magnus Pfeffer im Rahmen der European Conference on Data Analysis 2013, http://digbib.ubka.uni-karlsruhe.de/volltexte/1000035777

PPN:	611235307
Titel:	Zwanghafte Persönlichkeitsstörung und Zwangserkrankungen : Therapie und Selbsthilfe / Nicolas Hoffmann; Birgit Hofmann
Verfasser:	Hoffmann, Nicolas *1940-* ; Hofmann, Birgit *1965-*
Ort/Jahr:	Berlin [u.a.] : Springer, 2010
Umfang:	XII, 154 S. : Ill. ; 25 cm
Anmerkung:	Literaturverz. S. 147
ISBN:	978-3-642-02513-6*Pp. : EUR 34.95, ca. sfr 51.00 (freier Pr.)

PPN:	626919819
	[Elektronische Ressource]
Titel:	Zwanghafte Persönlichkeitsstörung und Zwangserkrankungen : Therapie und Selbsthilfe / von Nicolas Hoffmann, Birgit Hofmann
Verfasser:	Hoffmann, Nicolas
Sonst. Personen:	Hofmann, Birgit
Ort/Jahr:	Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2010
Umfang:	Online-Ressource. : v.: digital.
Andere Ausgabe:	Print version: Zwanghafte Persönlichkeitsstörung und Zwangserkrankungen: Therapie und Selbsthilfe (German Edition)
Anmerkung:	Includes bibliographical references and index
ISBN:	978-3-642-02514-3 Weitere Ausgaben: 978-3-642-02513-6 (Printausgabe)

Abbildung 2: Zwei Ausgaben desselben Werkes im GBV – einmal mit und einmal ohne Individualisierung

Anwendung des Verfahrens auf die Daten aus allen Katalogen.[2]

Automatisches Generieren von Konkordanzen

Die gefundenen Werk-Cluster sind durch die aggregierten Erschließungsinformationen eine gute Ausgangsbasis für weitere Auswertungen. So lassen sich zum Beispiel aus dem gemeinsamen Auftreten von Notationen oder Schlagwörtern aus unterschiedlichen Erschließungssystemen in Clustern Zusammenhänge erkennen.[3] Für zwei unterschiedliche Erschließungssysteme – zum Beispiel die Dewey Decimal Classification (DDC) und die RVK – werden dazu nur die Cluster betrachtet, die Notationen aus beiden Systemen enthalten. Für jede einzelne Notation der RVK wird im nächsten Schritt ermittelt, in wie vielen Clustern sie insgesamt auftritt und wie häufig davon gemeinsam mit einer oder mehreren Notationen der DDC. Dabei wird für jedes Paar aufsummiert, wie häufig es gemeinsam auftritt.

Aus dem Verhältnis des alleinigen zum gemeinsamen Auftreten sowie der Anzahl der »Partner«, welche eine Notation besitzt, lassen sich Rückschlüsse ziehen: Findet sich für eine RVK-Notation nur eine gemeinsam auftretende DDC-Notation, so impliziert dies eine enge Übereinstimmung zwischen den beiden. Tritt eine Notation der RVK mit mehreren Notationen der DDC auf, impliziert dies, dass die gefundenen Notationen der DDC zusammen einer Notation der RVK entsprechen – die RVK-Notation ist also übergeordnet. Die so gewonnenen Beziehungen lassen sich zu einer automatisch generierten Konkordanz aggregieren.

Dieses Verfahren ist auch als »Instance Matching« bekannt und wurde bereits erfolgreich auf bibliografische Daten angewendet: In einem Projekt wurde eine Konkordanz zwischen dem Thesaurus und der Klassifikation der Niederländischen Nationalbibliothek erstellt und anschließend von Kollegen der Königlichen Bibliothek in Den Haag begutachtet und für gut befunden.[4] Für die Anwendung auf die heterogenen deutschen Katalogdaten ist ein vorheriges Clustering allerdings sehr wichtig: Durch das Zusammenführen entsteht idealerweise für jedes Werk genau ein Cluster, in dem alle Erschließungsinformationen gebündelt sind. Dadurch stehen zum einen mehr mögliche Notationspaare zur Verfügung. Zum anderen ist sichergestellt, dass doppelt vorhandene Einträge oder eine hohe Zahl von Ausgaben eines Titels nicht dazu führen, dass die mit ihm verbundenen Notationspaare zu stark gewichtet werden.

Diese sogenannte »Individualisierung« ist bisher nur unvollkommen umgesetzt: Zum Teil reichen die vorliegenden Informationen nicht aus oder der Aufwand ist zu hoch.

Individualisierung von Personen

Das Werk-Clustering bietet jedoch noch weitaus mehr Möglichkeiten.[5] Denn nicht nur die Sacherschließung, sondern auch andere Informationen sind konzeptionell auf der Ebene des Werkes angesiedelt und können deshalb von der Methode profitieren. Ein besonders interessantes Feld stellen dabei die Personennormsätze dar.

Seit längerer Zeit werden bei Personen zusätzliche Informationen wie zum Beispiel Lebensdaten, Beruf und Geburtsort erfasst. Dies ermöglicht es beispielsweise, den Schriftsteller Heiner Müller von einem gleichnamigen, 1970 in Erfurt geborenen Arzt sowie weiteren Personen mit demselben Namen zu unterscheiden. Diese sogenannte »Individualisierung« ist bisher jedoch nur unvollkommen umgesetzt: Zum Teil reichen die vorliegenden Informationen für eine Individualisierung nicht aus oder der Aufwand dafür ist zu hoch. Oft ist auch die Zuordnung neuer Titeldatensätze zu den vorhandenen Personennormsätzen schwierig und fehleranfällig. Ein besonderes Problem stellen überdies die nur zum Teil aufgearbeiteten Altdaten sowie maschinell eingespielte Verlagsdaten (zum Beispiel bei E-Book-Paketen) dar.

Hier bietet das Werk-Clustering große Chancen: Theoretisch genügt es, wenn in einem einzigen Katalog ein einziger

4 Vergleiche A. Isaac, L. van der Meij, S. Schlobach, S. Wang (2007): An empirical study of instance-based ontology matching. In: The semantic web: 6th International Semantic Web Conference; proceedings, Berlin, Heidelberg: Springer, Seite 253–266

5 Vgl. dazu die Präsentation von Heidrun Wiesenmüller auf dem Workshop, URL: http://de.slideshare.net/heidrunw/wiesenmueller-bibliothekskongress-2013workshopclustering

6 Im vorliegenden Beispiel ergab eine Recherche, dass nicht nur eine, sondern sogar beide Zuordnungen falsch sind: Beim tatsächlichen Autor handelt es sich um einen 1967 in Roßlau geborenen Kunsthistoriker und Publizisten.

7 Ein noch größerer Erfolg wäre zu erwarten, wenn neben deutschen auch angloamerikanische Katalogdaten in den Abgleich mit einbezogen würden. Denn in der angloamerikanischen Welt ist die Individualisierung schon seit Langem normale Praxis.

Titeldatensatz aus einem Werk-Cluster mit dem richtigen individualisierten Personennormsatz verknüpft ist – dann kann diese Zuordnung auch für alle anderen Mitglieder des Clusters übernommen werden. Häufig ist beispielsweise derselbe Titel in einem Teil der Verbünde mit einem individualisierten Personensatz verknüpft, in anderen Verbünden hingegen nur mit einem Namenssatz. Dieser enthält nur den Namen und kann für verschiedene gleichnamige Personen verwendet werden. Sogar im selben Katalog kann es zu einem Nebeneinander der beiden Typen kommen. Abbildung 2 (Seite 626) zeigt zwei Ausgaben desselben Werkes im Gemeinsamen Bibliotheksverbund (GBV): Die Druckausgabe ist mit den individualisierten Normsätzen der beiden Verfasser verknüpft, die E-Book-Ausgabe hingegen nur mit Namenssätzen.

Auch Zuordnungsfehler können mit der Clustering-Methode erkannt werden.

Auch das Auflösen von sogenannten »Sammeltöpfen« könnte durch das Werk-Clustering unterstützt werden. Gemeint sind damit Normsätze für gängige Namen, die oft mit einer großen Zahl von Titeln unterschiedlicher Autoren verknüpft sind. Im SWB gibt es beispielsweise einen Namenssatz »Müller, Peter« mit über sechshundert zugehörigen Titeln. Bisher kann ein solcher Sammeltopf nur intellektuell in mühevoller Kleinarbeit auseinander gezogen werden. Über das Werk-Clustering könnte man wahrscheinlich einen großen Teil der Titel maschinell den richtigen Personensätzen zuweisen.

Auch Zuordnungsfehler können mit der Clustering-Methode erkannt werden. Abbildung 3 (diese Seite) zeigt das Werk »Symbol mit Aussicht« eines Autors namens Peter Müller in zwei Verbundkatalogen. Im SWB ist der Titel mit einem 1947 geborenen Kunsthistoriker verknüpft, im hbz mit einem 1936 geborenen Kunsthistoriker. Dass nicht beides gleichzeitig richtig sein kann, liegt auf der Hand. In solchen Fällen könnte eine Warnung ausgegeben werden, damit der Fall intellektuell überprüft werden kann.[6]

Sowohl Quantität als auch Qualität der Individualisierung kann also durch das Werk-Clustering erheblich gesteigert werden,[7] was wiederum neue Funktionalitäten in unseren Katalogen ermöglicht. Bisher nämlich kommen die mit hohem Aufwand erstellten Individualisierungs-

PPN:	065006429 Zitier
Titel:	Symbol mit Aussicht : die Geschichte des Berliner Fernsehturms / Peter Müller
Verfasser:	Müller, Peter [1947-]
Erschienen:	Berlin : Verl. Bauwesen, 1999
Umfang:	176 S. : Ill.
ISBN:	3-345-00651-0

1. Person	Müller, Peter, 1936- [(DE-588)123914078]
Titel	Symbol mit Aussicht
Untertitel	die Geschichte des Berliner Fernsehturms
Verfasser/Urheber	Peter Müller
Ort	Berlin
Verlag	Verl. für Bauwesen
Jahr	1999
Umfangsang.	176 S. : Ill., Kt.
ISBN	3-345-00651-0 Pp. DM 49.80

Abbildung 3: Verknüpfung desselben Werkes mit unterschiedlichen Personensätzen

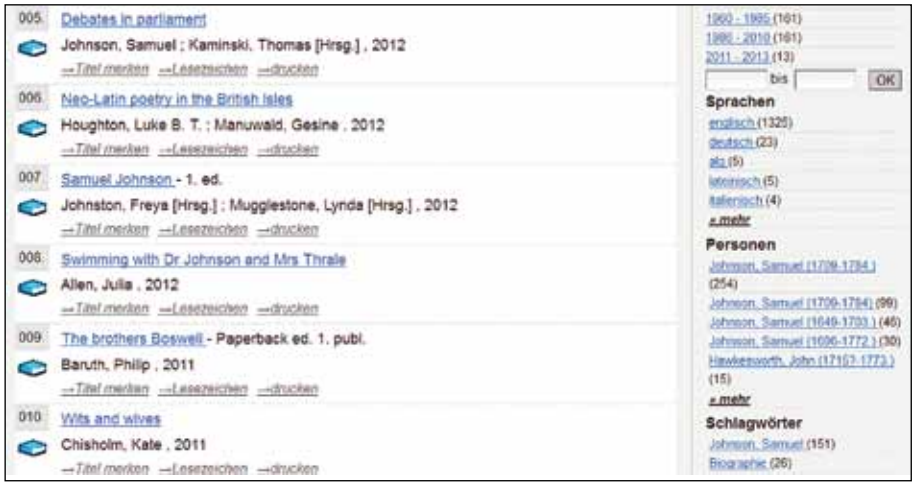

Abbildung 4: Drill-down mit individualisierten Personen im Freiburger Katalog plus maschinell erstellte Normsätze für Werke

Abbildung 5: Zusammenführung von Ausgaben eines Werkes im Primo-Katalog der UB Mannheim

daten unseren Nutzern kaum zugute – in den Trefferlisten werden üblicherweise die Titel aller gleichnamigen Personen einfach zusammengeworfen. Dass es auch anders geht, zeigt der Katalog plus der UB Freiburg[8]: Hier können die Titel namensgleicher Autoren über eine Drill-down-Facette unterschieden werden (Abbildung 4, Seite 627).

Ein weiteres zukunftsträchtiges Anwendungsfeld für das Clustering ist die »FRBRisierung« von Bibliothekskatalogen. Zentral ist dabei das Konzept des Werkes: Alle Manifestationen (Ausgaben) eines Werkes sollen im Katalog übersichtlich zusammengeführt und den Nutzern zur Auswahl präsentiert werden. Einige Katalogsysteme (zum Beispiel Primo von Ex Libris) realisieren dies über einen eingebauten Clustering-Mechanismus. Abbildung 5 (Seite 627) zeigt die entsprechende Funktion im Primo-Katalog der UB Mannheim.[9] Die Information ist dabei nicht fest in den Daten »verdrahtet«, sondern wird »on the fly« über einen entsprechenden Algorithmus erstellt. Der Nachteil einer solchen Lösung ist, dass sie nur im jeweiligen System funktioniert und in anderen Katalogen nicht nachgenutzt werden kann.

> Im Bereich der Musik wird eine weitaus elegantere Methode praktiziert: Für jedes musikalische Werk wird ein Normsatz erstellt, welcher mit allen zugehörigen Titeldatensätzen verknüpft wird.

Eine datentechnische Verknüpfung verschiedener Ausgaben gibt es bisher nur in bestimmten Fällen – beispielsweise werden parallele Druck- und Online-Ausgaben über Fußnoten miteinander verlinkt. Im Bereich der Musik wird hingegen eine weitaus elegantere Methode praktiziert: Für jedes musikalische Werk wird ein Normsatz erstellt, welcher mit allen zugehörigen Titeldatensätzen verknüpft wird. Abbildung 6 (diese Seite) zeigt den Werknormsatz für Mozarts Oper »Così fan tutte« in der Gemeinsamen Normdatei (GND). Eine entsprechende Lösung wäre auch für andere Werke wünschenswert – allerdings viel zu aufwendig, wenn die Werknormsätze manuell erstellt und verknüpft werden müssen.

Auch hier wäre eine Lösung über das Werk-Clustering denkbar, indem man aus den Titeldatensätzen eines Clusters die werkrelevanten Informationen extrahiert. Beispielsweise könnte das Jahr der frühesten vorhandenen Ausgabe als mutmaßliches Jahr des Werkes gelten. Auch die Originalsprache des Werkes lässt sich mit einiger Wahrscheinlichkeit ermitteln. Der Einheitssachtitel wäre als Titel des Werkes aufzufassen, Übersetzungstitel würden als Verweisungen gespeichert. Gemäß entsprechender Ableitungsregeln würde aus den vorhandenen Informationen maschinell ein Werknormsatz erstellt und mit allen Mitgliedern des Clusters verknüpft.

Vor dem Hintergrund des Umstiegs auf den neuen Katalogisierungsstandard »Resource Description and Access« (RDA) ist diese Aussicht besonders reizvoll. Denn RDA basiert auf dem FRBR-Modell und strebt deshalb auch die Abbildung der Beziehungen zwischen einem Werk, seinen Expressionen (Realisierungen) und Manifestationen an. Zwar stellt RDA an diese Abbildung keine hohen Ansprüche, doch ist klar, dass man die Vorteile des neuen Regelwerks umso stärker nützen kann, je besser das FRBR-Modell in den Daten umgesetzt wird. Eine durchgängige Verknüpfung unserer Titeldaten mit Werknormsätzen wäre ein großer Schritt nach vorne. Auch die DNB ist deshalb sehr daran interessiert, ein maschinelles Verfahren für Werknormsätze zu entwickeln.

Allerdings gibt es noch viele offene Fragen: Wie sollen Werke behandelt werden, von denen es nur eine einzige Ausgabe gibt? Was passiert, wenn bereits ein intellektuell erstellter Werknormsatz aus der Sacherschließung (das heißt ein Schlagwort für das Werk) vorhanden ist? Sollen maschinell erstellte Werknormsätze in die GND eingespielt werden? Und natürlich: Lässt sich das geschilderte Szenario unter den derzeitigen technischen Rahmenbedingungen mit mehreren Verbunddatenbanken überhaupt umsetzen?

GND

Link zu diesem Datensatz	http://d-nb.info/gnd/300107331
Komponist/Urheber	Mozart, Wolfgang Amadeus
Titel des Werkes	Così fan tutte
Andere Titel	La scuola degli amanti Così fan tutte ossia la scuola degli amanti So treiben's alle
Quelle	KV
Erläuterungen	Definition: Dramma giocoso in 2 Akten Verwendungshinweis: Ansetzung nach den RAK-M 2003
Zeit	erstellt: 1790
Land	Österreich (XA-AT)
Oberbegriffe	Beispiel für: Dramma giocoso Beispiel für: Oper
Systematik	14.4p Personen zu Musik
Typ	Werk der Musik (wim)

Abbildung 6: Normdatensatz für ein musikalisches Werk in der GND

Optimierungspotenziale für das Clustering-Verfahren

Bei der Erstellung der Werk-Cluster gibt es noch Verbesserungsmöglichkeiten. Dabei muss je nach Anwendungsfall entschieden werden, wie »scharf« das Clustering eingestellt werden soll. Die bisher angewendete Methode ist daraufhin optimiert, Fehl-Zusammenführungen zu vermeiden. Sie sorgt also dafür, dass sich kein Fremdkörper (in Form einer Ausgabe eines anderen Werkes) in das Cluster einschleicht. Erreicht wird dies durch einen sehr strengen Abgleich: Zusätzlich zu einer Person oder Körperschaft müssen, wenn es keinen Einheitssachtitel gibt, nicht nur der Sachtitel, sondern auch die Titelzusätze exakt übereinstimmen. Gerade bei Zusätzen gibt es aber nicht selten kleine Unterschiede zwischen den Ausgaben. Entsprechend wird bei einem scharfen Clustering nicht alles das, was zusammengehört, auch tatsächlich zusammengeführt.

Professorin Heidrun Wiesenmüller M.A., geboren 1968 in Nürnberg, studierte Mittlere Geschichte, Anglistik und Mittellatein in Erlangen und Newcastle upon Tyne. Nach dem Referendariat an der Landesbibliothek Oldenburg und der FH Köln war sie zunächst als Fachreferentin an der Württembergischen Landesbibliothek tätig. Seit 2006 lehrt sie Formal- und Sacherschließung an der Hochschule der Medien in Stuttgart. Sie ist Mitglied verschiedener regionaler und überregionaler Fachgremien, darunter auch der AG RDA des Standardisierungsausschusses. – Kontakt: wiesenmueller@hdm-stuttgart.de

Für manche Anwendungen ist dieser Ansatz dennoch sinnvoll: Für die Übernahme von Sacherschließung beispielsweise ist es unproblematisch, wenn in manchen Fällen ein Werk in mehrere Cluster zerfällt. Für die maschinelle Erstellung von Werknormsätzen hingegen ist eine möglichst große Vollständigkeit innerhalb des Clusters erwünscht. Entsprechend muss ein gewisser Unschärfe-Faktor in den Algorithmus eingebaut werden. In seltenen Fällen wird dies dazu führen, dass ein falscher Titel ins Cluster eingeordnet wird – solche Fehler müssen wohl oder übel in Kauf genommen werden.

> Mit den Datenmengen aller deutschsprachigen Verbundkataloge und weiterer externer Quellen werden Metadaten im Bereich von 100 Millionen Einträgen gesammelt.

Durch den Einbezug weiterer Informationen kann die Zusammenführung noch verbessert werden. So werden bisher bei Personen und Körperschaften nur die Ansetzungsformen miteinander verglichen – es sollten jedoch auch Verweisungsformen berücksichtigt werden. Ebenfalls noch nicht ausgewertet werden bislang Fußnoten, in denen auf eine Titeländerung hingewiesen wird (»2. Aufl. u.d.T.: …«).

Metafacture als neue technische Plattform

Eine grundsätzliche Herausforderung der vorgestellten und auch anderer denkbarer Verfahren zur Bearbeitung von Metadaten ist der dazu erforderliche Speicher- und Rechenzeitaufwand. Solange nur kleine Datenmengen verarbeitet werden, ist dies noch nicht problematisch. Aber mit den Datenmengen aller deutschsprachigen Verbundkataloge und weiterer externer Quellen werden Metadaten im Bereich von 100 Millionen Einträgen gesammelt. Diese können nicht mehr auf einem einzelnen Rechner verarbeitet werden – hier ist eine Verteilung auf mehrere Rechner erforderlich. Dazu kommt, dass alle in den Projektphasen entwickelten Programme als wenig dokumentierte Prototypen mithilfe einer Skriptsprache entstanden sind. Für einen dauerhaften produktiven Einsatz durch Dritte sind sie nicht geeignet. Benötigt wird vielmehr eine solide technische Basis, die in einer verbreiteten und plattformunabhängigen Programmiersprache geschrieben und vollständig dokumentiert ist. Nur so kann eine gemeinschaftliche Weiterentwicklung der Verfahren stattfinden, und Interessierte könnten darauf aufbauend eigene Anpassungen vornehmen.

Das hbz und die DNB standen vor einem ähnlichen Problem und entwickelten eine solche Basis im Rahmen des gemeinsamen Projekts »culturegraph« zum Aufbau eines Resolving- und Lookup-Dienstes für bibliothekarische Identifier.

> Die eigene Erschließungsarbeit wird zwar nicht verschwinden, künftig jedoch einen geringeren Raum einnehmen.

Die in der Programmiersprache Java geschriebene Softwareplattform »Metafacture« bündelt die modular aufgebauten Programme, stützt sich auf bereits existierende freie Software, wie Apache Hadoop, HBase und Lucene, und ist gut dokumentiert. Durch den Einsatz von Hadoop, einem Framework zur Entwicklung von skalierbarer Software, lassen sich die Verarbeitungsschritte auf mehrere Rechner verteilen.

Um Anpassungen und Erweiterungen weiter zu erleichtern, wurden für die Softwareplattform eigene Beschreibungssprachen entwickelt: »MetaMorph« stellt Funktionen bereit, die typischerweise bei der Verarbeitung von bibliografischen Daten benötigt werden; »MetaFlow« erlaubt es, mehrere Verarbeitungs- und Auswertungsschritte zu verketten.[10] Sie haben das langfristige Potenzial dafür, dass auch Nicht-Programmierer eigene Ideen mit Metadaten umsetzen können.

Professor Magnus Pfeffer ist Jahrgang 1974 und studierte Informatik an der Universität Kaiserslautern. Er war ab 2003 Fachreferent an der Universitätbibliothek Mannheim und dort für zahlreiche IT-Projekte verantwortlich. Seit Herbst 2011 lehrt er an der Hochschule der Medien in Stuttgart. Seine Interessenschwerpunkte sind unter anderem automatische Erschließung und Anwendungen von Technologien des Semantic Web in Bibliotheken. – Kontakt: pfeffer@hdm-stuttgart.de

Ausblick

Alle im Workshop vorgestellten Verfahren und Analysen möchte Magnus Pfeffer nach und nach in Metafacture integrieren. Sie werden dann – wie die gesamte Plattform – unter einer freien und quelloffenen Lizenz jedermann zur Verfügung stehen. Die DNB unterstützt ihn dabei in einem gemeinsamen Projekt.

Bibliothekare müssen – davon sind die Organisatoren des Workshops überzeugt – verstärkt zu »Metadaten-Managern« werden. Die eigene Erschließungsarbeit wird zwar nicht verschwinden, künftig jedoch einen geringeren Raum einnehmen. Unser Know-how wird stattdessen gefragt sein, um immer heterogener werdende Daten zu vereinheitlichen, zu verbessern und aufzuwerten. Dafür ist nicht nur ein neues Selbstverständnis nötig, sondern es muss auch neue technische Methoden und Werkzeuge geben. Ein Beispiel dafür bietet das beschriebene Clustering-Verfahren, dessen Potenziale noch längst nicht ausgeschöpft sind.

8 http://katalog.ub.uni-freiburg.de/opac. Die genannte Drill-down-Möglichkeit besteht nur im Reiter »Bücher & mehr«.

9 www.bib.uni-mannheim.de/133.html

10 Beide wurden im Workshop von Markus Geipel (DNB) vorgestellt und mit Beispielen illustriert. Die Software und Dokumentation findet sich auf https://github.com/culturegraph.